

Максимизация числа допустимых ошибок аппроксимации линейной регрессии с детерминированными ограничениями

С.И. Носков, В.М. Лацвиев

Аннотация — В работе рассматривается задача построения линейной регрессионной модели, направленная на максимизацию числа наблюдений, для которых ошибка аппроксимации не превышает заданного порогового значения. В отличие от классических методов, минимизирующих сумму квадратов или модулей ошибок, предложенный подход позволяет непосредственно управлять долей допустимых отклонений. Эта задача формулируется в виде задачи линейно-булевого программирования (ЛБП). В дополнение к исходной модели предложено шесть типов детерминированных ограничений, отражающих априорные экспертные знания о предметной области: интервальные ограничения на параметры, ограничения на сумму и максимум абсолютных ошибок, знаковая совместимость коэффициентов, требование точных попаданий (наличия заданного числа наблюдений с нулевой ошибкой) и равномерность вклада признаков по наблюдениям. Показано, что все введенные ограничения допускают линеаризацию, что сохраняет принадлежность задачи к классу ЛБП. Сформулировано и доказано утверждение об эквивалентности исходной и линеаризованной постановок. Предложенный подход применён для моделирования пассажирооборота воздушного транспорта Российской Федерации на основе статистических данных за 2002–2019 годы. Сравнительный анализ с ранее опубликованными результатами показывает, что введение дополнительных ограничений приводит к ужесточению допустимого множества решений: число допустимых ошибок при этом уменьшается с 11 до 8, однако сумма модулей ошибок снижается со 175.6 до 167.8, что свидетельствует о повышении точности аппроксимации на всём множестве наблюдений.

Ключевые слова — регрессия, линейно-булево программирование, априорная информация, детерминированные ограничения.

I. ВВЕДЕНИЕ

Построение регрессионных моделей, учитывающих априорные ограничения, формализующие некоторые особенности функционирования объекта исследований, является одной из важных задач анализа данных. Классический метод наименьших квадратов (МНК) в своей стандартной формулировке не предназначен для непосредственного включения таких ограничений, а их добавление превращает задачу в более сложный случай

условной оптимизации. Например, в эконометрике хорошо известны модели, где сумма коэффициентов регрессии должна равняться единице. Для оценки параметров в таких условиях применяются специальные процедуры [1]. Однако подобные жёсткие условия не всегда отражают реальные ситуации, когда знание приобретает форму неравенств или диапазонов. Работы [2,3] развили теорию регрессионного оценивания с ограничениями-неравенствами, показав, что их учёт повышает точность и интерпретируемость оценок.

Особый класс составляют знаковые ограничения, когда из физических или экономических соображений заранее известно, что влияние фактора не может быть отрицательным. Исследования [4] показали, что для широкого класса задач неотрицательная регрессия устойчива к переобучению и может быть столь же эффективна, как и L_1 -регуляризация. Позднее этот подход был распространён на метод LASSO с ограничениями на знаки [5].

В инженерных и естественно-научных задачах особенно важны ограничения, вытекающие из фундаментальных законов природы. Работы по регрессии с ограничениями на форму зависимости (монотонность, выпуклость) показывают, что такие модели обладают предсказуемым поведением даже за пределами обучающей выборки [6]. Эти ограничения называют «жёсткими» (hard constraints) [7]. Параллельно развивались байесовские методы работы с ограничениями-неравенствами и неопределённостью в спецификации ограничений [8].

Все перечисленные методы объединяет одно: они нацелены на нахождение единственного вектора коэффициентов, который минимизирует сумму квадратов (или модулей) ошибок при заданных ограничениях. Однако в некоторых прикладных областях гораздо важнее не минимизировать среднюю ошибку, а максимизировать число наблюдений, для которых ошибка не превышает заданного порога. Эта идея была сформулирована в работе [9], где задача сведена к линейно-булеву программированию (ЛБП).

В работе [10] предложено использовать смешанно-целочисленную оптимизацию для задачи выбора наилучшего подмножества регрессоров, что также позволяет учитывать ограничения на коэффициенты.

Таким образом, возникает необходимость объединить два направления: учёт разнообразных априорных ограничений (знаковых, интервальных, физических) и

Статья получена 05 июня 2026.

С.И. Носков, ФГБОУ ВО ИрГУПС, профессор, д.т.н. (e-mail: sergey.noskov.57@mail.ru)

В.М. Лацвиев, ФГБОУ ВО ИрГУПС, аспирант (e-mail: slava.latsviev.02@mail.ru)

альтернативный критерий качества – максимизацию числа наблюдений, попавших в допустимый интервал ошибки. В данной работе предлагается такое объединение для шести типов детерминированных ограничений.

II. ПОСТАНОВКА И ФОРМАЛИЗАЦИЯ ЗАДАЧИ

Рассмотрим линейную регрессионную модель:

$$y_k = \sum_{i=1}^m \alpha_i x_{ki} + \varepsilon_k, k = \overline{1, n}. \quad (1)$$

Здесь k – номер наблюдения, y_k – зависимая переменная, x_{ki} – i -я независимая переменная (детерминированные величины), $\alpha_i, i = \overline{1, m}$ – неизвестные параметры, $\varepsilon_k, k = \overline{1, n}$ – ошибки аппроксимации.

В отличие от традиционного для регрессионного анализа подхода, связанного с минимизацией выбранной функции потерь при оценивании неизвестных параметров модели (1) (см, например, [1], [11]), будем применять для этого способ, связанный с максимизацией числа допустимых ошибок [9], для которых $|\varepsilon_k| \leq d$, где d – некий заранее заданный порог.

Введём неотрицательные переменные $u_k \geq 0, v_k \geq 0, k = \overline{1, n}$ такие, что:

$$\varepsilon_k = u_k - v_k, |\varepsilon_k| = u_k + v_k, u_k v_k = 0. \quad (2)$$

Тогда модель (1) представима в виде:

$$\sum_{i=1}^m \alpha_i x_{ki} + u_k - v_k = y_k, k = \overline{1, n}. \quad (3)$$

Булевы переменные $\sigma_k \in \{0, 1\}$ введем по правилу:

$$\sigma_k = \begin{cases} 1, & |\varepsilon_k| \leq d \\ 0, & |\varepsilon_k| > d. \end{cases} \quad (4)$$

Связь между u_k, v_k и σ_k задаётся с помощью наперед заданной большой положительной константы M :

$$u_k + v_k \leq d + M(1 - \sigma_k), k = \overline{1, n}. \quad (5)$$

Целевая функция, таким образом, имеет вид:

$$\sum_{k=1}^n \sigma_k - \delta \sum_{k=1}^n (u_k + v_k) \rightarrow \max, \quad (6)$$

где $\delta > 0$ – малое число, обеспечивающее выполнение условия $u_k v_k = 0$ (штраф за одновременную положительность u_k и v_k).

Введем далее шесть типов дополнительных детерминированных ограничений.

2.1. Интервальные ограничения на параметры

Пусть из априорных соображений известны границы для каждого параметра:

$$\alpha_i^{\min} \leq \alpha_i \leq \alpha_i^{\max}, i = \overline{1, m}. \quad (7)$$

2.2. Ограничение на суммарную абсолютную ошибку

Ограничим сумму модулей ошибок:

$$\sum_{k=1}^n (u_k + v_k) \leq E_{\max}. \quad (8)$$

2.3. Ограничение на максимальную абсолютную ошибку

Предусмотрим предотвращение больших ошибок на отдельных наблюдениях:

$$u_k + v_k \leq R, k = \overline{1, n}, \quad (9)$$

где R – заранее заданное положительное число.

2.4. Знаковая совместимость параметров

Пусть для заданного множества пар индексов $H \subseteq \{1, \dots, m\} \times \{1, \dots, m\}$ требуется, чтобы соответствующие параметры имели одинаковый знак: $\alpha_i \cdot \alpha_j \geq 0$ для всех $(i, j) \in H$. Для линейаризации этого условия введём булевы переменные $z_{ij} \in \{0, 1\}$ для каждой пары $(i, j) \in H$ и ограничения:

$$\begin{cases} \alpha_i \leq M z_{ij}, \\ -\alpha_i \leq M(1 - z_{ij}), \\ \alpha_j \leq M z_{ij}, \\ -\alpha_j \leq M(1 - z_{ij}). \end{cases} \quad (i, j) \in H. \quad (10)$$

Если $z_{ij} = 1$, то из первого и третьего неравенств имеем $\alpha_i \leq M, \alpha_j \leq M$, а из второго и четвёртого: $-\alpha_i \leq 0, -\alpha_j \leq 0$, то есть $\alpha_i \geq 0, \alpha_j \geq 0$. Если $z_{ij} = 0$, то из второго и четвёртого: $-\alpha_i \leq M, -\alpha_j \leq M$ (всегда выполняется при достаточно большом M), а из первого и третьего: $\alpha_i \leq 0, \alpha_j \leq 0$. Таким образом, при любом z_{ij} знаки α_i и α_j совпадают (оба неотрицательны или оба неположительны).

2.5. Требование точных попаданий

Пусть необходимо, чтобы не менее N_0 наблюдений были аппроксимированы точно, с нулевой ошибкой. Введём булевы переменные $\tau_k \in \{0, 1\}$ как индикаторы точности:

$$u_k + v_k \leq M(1 - \tau_k), \sum_{k=1}^n \tau_k \geq N_0. \quad (11)$$

При $\tau_k = 1$ имеем $u_k + v_k \leq 0$, откуда с учётом неотрицательности u_k, v_k получаем $u_k = v_k = 0$, что обеспечивает $\varepsilon_k = 0$.

2.6. Равномерность вклада признаков

Вклад каждого i -го признака в правую часть модели (введен в [11]) не должен слишком сильно варьироваться по наблюдениям:

$$\max_{k=\overline{1, n}} x_{ki} \alpha_i - \min_{k=\overline{1, n}} x_{ki} \alpha_i \leq U_i, i = \overline{1, m}, \quad (12)$$

где положительные числа $U_i, i = \overline{1, m}$ заранее заданы. Для линейаризации условий (12) достаточно попарных сравнений. Раскрывая разность, получаем для каждой тройки i, k, l два линейных неравенства:

$$\begin{aligned} x_{ki} \alpha_i - x_{li} \alpha_i &\leq U_i, x_{li} \alpha_i - x_{ki} \alpha_i \leq U_i, \\ i &= \overline{1, m}, k, l = \overline{1, n}. \end{aligned} \quad (13)$$

Эти неравенства эквивалентны $|x_{ki} \alpha_i - x_{li} \alpha_i| \leq U_i$, что в свою очередь влечёт выполнение (12). Обратное, если (12) выполнено, то для любых k, l разность не превосходит U_i , поэтому (13) также выполнено. Таким образом, (13) – точная линейная переформулировка (12).

Совместное использование исходной модели (3)–(6) и любого подмножества из описанных выше ограничений (7)–(13) приводит к задаче, остающейся в рамках линейно-булева программирования. Все ограничения линейны, а булевы переменные σ_k, τ_k, z_{ij} обеспечивают корректную обработку разрывных условий (допустимая ошибка, точное попадание, знаковая совместимость).

III. ОБЪЕДИНЁННАЯ ФОРМУЛИРОВКА ЗАДАЧИ ЛБП

Таким образом, объединённая задача ЛБП, реализующая сформулированную выше задачу оценивания параметров модели (1), включает целевую функцию (6) и ограничения (3), (5), (7), (8), (9), (10), (11), (13).

Все переменные удовлетворяют условиям:

$$u_k \geq 0, v_k \geq 0, \sigma_k \in \{0,1\}, \tau_k \in \{0,1\}, \quad k = \overline{1, n}, \quad (14)$$

$$z_{ij} \in \{0,1\}, (i,j) \in H.$$

Все ограничения линейны относительно непрерывных переменных α_i, u_k, v_k и булевых переменных σ_k, τ_k, z_{ij} . Общая размерность задачи составляет: количество непрерывных переменных $m + 2n$, количество булевых переменных $2n + |S|$ ($|S|$ – мощность множества S), количество ограничений $4n + 2m + 4|S| + 2 + 2mn^2$.

Отметим, что квадратичная зависимость числа ограничений от количества наблюдений ($2mn^2$) в случае больших выборок может существенно увеличить вычислительную сложность задачи ЛБП. Для повышения практической применимости алгоритма и сокращения числа ограничений (13) возможен переход от полного попарного сравнения к локальному. Например, для динамических рядов экономически обоснованно контролировать равномерность вклада признаков только для хронологически соседних наблюдений, что снижает число таких неравенств до $2m(n-1)$. Другим эффективным подходом является предварительная сортировка выборки по значениям факторов, позволяющая ограничить область попарных сравнений лишь близкими по значению наблюдениями. Оба способа сохраняют линейность ограничений и, следовательно, принадлежность задачи к классу ЛБП.

Некоторые свойства оптимального решения задачи ЛБП (6), (3), (5), (7), (8), (9), (10), (11), (13), (14) отражены в следующем утверждении.

Утверждение. Пусть $(\alpha^*, u^*, v^*, \sigma^*, \tau^*, z^*)$ – оптимальное решение задачи ЛБП (6), (3), (5), (7), (8), (9), (10), (11), (13), (14). Тогда:

1. α^* является вектором параметров линейной регрессионной модели, максимизирующей число наблюдений с ошибкой, не превышающей d , при выполнении всех заданных дополнительных ограничений (7)–(13).

2. При достаточно малом $\delta > 0$ в оптимальном решении выполняется условие $u_k v_k = 0$ для всех $k = \overline{1, n}$, следовательно, $|\varepsilon_k| = u_k + v_k$.

Доказательство.

1. Покажем, что каждое из ограничений задачи ЛБП эквивалентно соответствующему условию исходной содержательной постановки.

Уравнение (3) прямым образом следует из (1) и (2) и является точным воспроизведением модели.

Ограничение (5): если $\sigma_k = 1$, то (5) принимает вид $u_k + v_k \leq d$, откуда с учётом (2) получаем $|\varepsilon_k| = u_k + v_k \leq d$. Если же $\sigma_k = 0$, то (5) даёт $u_k + v_k \leq d + M$, что при достаточно большом M всегда выполняется. Таким образом, (5) точно реализует определение (4).

Целевая функция (6): максимизация $\sum \sigma_k$ напрямую соответствует максимизации числа наблюдений с

$|\varepsilon_k| \leq d$. Член $-\delta \sum (u_k + v_k)$ добавлен для устранения неоднозначности, как будет показано в части 2.

Ограничения (7)–(9) являются прямой записью соответствующих требований.

Ограничение (10): при любом $z_{ij} \in \{0,1\}$ знаки α_i и α_j совпадают, что в точности соответствует условию $\alpha_i \cdot \alpha_j \geq 0$. Обратно: для любой пары α_i, α_j с одинаковыми знаками можно выбрать $z_{ij} = 1$, если они оба неотрицательны, и $z_{ij} = 0$, если они оба неположительны. При этом все неравенства (10) будут выполнены, если M достаточно велико. Таким образом, (10) эквивалентно знаковой совместимости.

Ограничение (11): если $\tau_k = 1$, то $u_k + v_k \leq 0$, откуда $u_k = v_k = 0$ и $\varepsilon_k = 0$. Сумма $\sum \tau_k \geq N_0$ гарантирует, что по крайней мере N_0 наблюдений имеют $\tau_k = 1$, а значит, и $\varepsilon_k = 0$. Если $\tau_k = 0$, ограничение $u_k + v_k \leq M$ не активно.

Ограничение (13): как показано в п. 2.6, система неравенств (13) эквивалентна условию (12), поскольку для любых k, l разность $|x_{ki}\alpha_i - x_{li}\alpha_i| \leq U_i$ тогда и только тогда, когда выполнены оба неравенства $x_{ki}\alpha_i - x_{li}\alpha_i \leq U_i$ и $x_{li}\alpha_i - x_{ki}\alpha_i \leq U_i$. В свою очередь, выполнение этих неравенств для всех пар k, l влечёт выполнение (12) и обратно.

2. Покажем теперь справедливость равенств $u_k v_k = 0$ для всех $k = \overline{1, n}$ при малом $\delta > 0$.

Предположим, что в некотором оптимальном решении нашёлся индекс k , для которого $u_k > 0$ и $v_k > 0$ одновременно. Рассмотрим два случая.

Случай 1: $\sigma_k = 1$. Тогда из (5) имеем $u_k + v_k \leq d$. Уменьшим одновременно u_k и v_k на величину $\varepsilon = \min(u_k, v_k) > 0$, то есть положим $u'_k = u_k - \varepsilon$, $v'_k = v_k - \varepsilon$. При этом:

– разность $u'_k - v'_k = (u_k - \varepsilon) - (v_k - \varepsilon) = u_k - v_k = \varepsilon_k$ не изменилась;

– сумма $u'_k + v'_k = (u_k + v_k) - 2\varepsilon$ уменьшилась на 2ε ;

– неравенство $u'_k + v'_k \leq d$ по-прежнему выполняется;

– остальные ограничения (7)–(13) не зависят от u_k, v_k или зависят только от их суммы, которая уменьшилась, поэтому они остаются выполненными;

– целевая функция изменится: $\sum \sigma_k$ останется прежним, а штрафной член $-\delta \sum (u_k + v_k)$ увеличится на $\delta \cdot 2\varepsilon > 0$. Следовательно, исходное решение не могло быть оптимальным – противоречие.

Случай 2: $\sigma_k = 0$. Тогда из (5) имеем $u_k + v_k \leq d + M$. Мы можем проделать ту же операцию уменьшения u_k и v_k на $\varepsilon = \min(u_k, v_k)$. При этом: ε_k не меняется; сумма $u_k + v_k$ уменьшается на 2ε , что увеличивает целевую функцию за счёт уменьшения штрафа; все ограничения, зависящие от $u_k + v_k$ (например, (8), (9), если они активны), остаются выполненными, так как сумма уменьшается. Снова получаем улучшение целевой функции, что противоречит оптимальности.

Следовательно, $u_k v_k = 0$ для всех k . В силу того, что $\varepsilon_k = u_k - v_k$ и $u_k, v_k \geq 0$, это влечёт $|\varepsilon_k| = u_k + v_k$.

Докажем оптимальность по числу допустимых ошибок.

Предположим, что существует другой вектор параметров α' , для которого число наблюдений с $|\varepsilon'_k| \leq d$ больше, чем у α^* и который удовлетворяет всем дополнительным ограничениям. Построим по нему допустимое решение задачи ЛБП: определим $\varepsilon'_k = y_k - \sum \alpha'_i x_{ki}$, положим $\sigma'_k = 1$, если $|\varepsilon'_k| \leq d$, иначе $\sigma'_k = 0$, выберем $u'_k = \max(\varepsilon'_k, 0)$, $v'_k = \max(-\varepsilon'_k, 0)$. Для ограничений (10) и (11) выберем соответствующие булевы переменные, удовлетворяющие условиям (существование таких переменных при достаточно большом M гарантировано, как показано выше). Тогда значение целевой функции для этого решения равно $\sum \sigma'_k - \delta \sum (u'_k + v'_k)$. При достаточно малом δ (например, $\delta < \frac{1}{\sum (u'_k + v'_k) + 1}$) первое слагаемое доминирует, и если $\sum \sigma'_k > \sum \sigma_k^*$, то значение целевой функции будет больше, чем у α^* , что противоречит оптимальности α^* в задаче ЛБП. Следовательно, α^* действительно максимизирует число наблюдений с $|\varepsilon_k| \leq d$ среди всех допустимых наборов параметров. Утверждение доказано.

IV. РЕГРЕССИОННАЯ МОДЕЛЬ ПАССАЖИРООБОРОТА ВОЗДУШНОГО ТРАНСПОРТА РОССИИ

Применим описанный способ построения линейной регрессии для моделирования пассажирооборота воздушного транспорта Российской Федерации с последующим сравнением с полученными в [9] результатами. Введем следующие обозначения: y – пассажирооборот воздушного транспорта, млрд. пасс. км; x_1 – среднемесячная заработная плата, руб.; x_2 – численность трудоспособного населения в России, млн. чел. В качестве информационной базы для моделирования используем статистическую ежегодную информацию за 2002–2019 гг. [12]. Будем строить линейную двухфакторную модель без свободного члена:

$$y_k = \sum_{i=1}^2 \alpha_i x_{ki} + \varepsilon_k, \quad k = \overline{1, 18}.$$

В [9] при $d=6$ путем решения задачи ЛБП (6), (3), (5) с частью соответствующих ограничений (14) получена модель

$$y_k = 0.00576x_{k1} + 0.42403x_{k2} + \varepsilon_k, \quad k = \overline{1, 18}, \quad (15)$$

$$\sum_{k=1}^{18} |\varepsilon_k| = 175.6, |S| = 11.$$

Зададим ограничения и необходимые константы:

$$\begin{aligned} d &= 6; \\ 0.006 &\leq \alpha_1 \leq 0.01, 0.3 \leq \alpha_2 \leq 0.4; \\ E_{\max} &= 200; R = 50; \\ \alpha_1 \cdot \alpha_2 &\geq 0; N_0 = 1; \\ U_1 &= 200, U_2 = 10; \\ \delta &= 0.0001; M = 10000. \end{aligned}$$

В результате решения задачи ЛБП (6), (3), (5), (7), (8), (9), (10), (11), (13), (14) с указанными числовыми параметрами получена следующая регрессионная модель:

$$y_k = 0.00608x_{k1} + 0.32528x_{k2} + \varepsilon_k, \quad k = \overline{1, 18}.$$

Сумма модулей ошибок аппроксимации составила $\sum_{k=1}^{18} |\varepsilon_k| = 167.8$, а число наблюдений, для которых

абсолютная ошибка не превышает порога $d = 6$, оказалось равным $|S| = 8$. Требование наличия хотя бы одного наблюдения с нулевой ошибкой ($N_0 = 1$) выполнено – таким наблюдением является $k = 5$ ($\varepsilon_5 = 0$).

Сравнение с моделью (15), полученной в [9] без учёта дополнительных ограничений показывает, что введение ограничений (8), (9), (12)–(13) привело к ужесточению допустимого множества решений. Как следствие, значение целевой функции (число допустимых ошибок) уменьшилось с 11 до 8, однако сумма модулей ошибок также снизилась со 175.6 до 167.8, что свидетельствует о повышении точности аппроксимации на всём множестве наблюдений. Полученные оценки параметров удовлетворили всем заданным интервальным и знаковым ограничениям, а также условию равномерности вклада признаков.

При необходимости исследователь может варьировать значения констант $E_{\max}$, R , U_i и N_0 , добиваясь приемлемого компромисса между максимизацией числа допустимых ошибок и другими требованиями к модели.

V. ЗАКЛЮЧЕНИЕ

1. Предложено расширение задачи максимизации числа допустимых ошибок аппроксимации за счёт включения шести типов дополнительных детерминированных ограничений.

2. Показано, что все ограничения сводятся к линейным неравенствам с булевыми переменными, что сохраняет принадлежность задачи к классу линейно-булева программирования.

3. Введены ограничения на сумму и максимум ошибок, попарные соотношения коэффициентов, требование точных попаданий и равномерность вклада признаков.

4. Предложенный подход позволяет учитывать содержательную априорную информацию без привлечения вероятностных гипотез.

5. При использовании предложенных способов сокращения числа ограничений (локальные сравнения или сортировка) задача сохраняет практическую разрешимость для выборок, характерных для большинства эконометрических исследований.

6. Выполненный численный эксперимент на данных пассажирооборота воздушного транспорта России показал, что введение дополнительных ограничений (8), (9), (12)–(13) при $d = 6$ позволило снизить сумму модулей ошибок с 175.6 до 167.8, однако привело к уменьшению числа допустимых ошибок с 11 до 8 по сравнению с моделью, описанной в [9]. Полученные оценки параметров удовлетворили всем заданным ограничениям.

7. Предложенный ЛБП-подход принципиально отличается от классических методов регуляризации (LASSO, гребневая регрессия) и байесовского оценивания. В то время как L_1 - и L_2 -регуляризация направлены на минимизацию суммы квадратов (или модулей) ошибок со штрафными членами, а байесовский подход [8] оперирует апостериорными распределениями параметров, предлагаемый критерий (6)

непосредственно ориентирован на управление числом наблюдений с ошибкой, не превышающей заданный порог. Это критично для прикладных задач, где требуется гарантировать качество аппроксимации для заданной доли объектов, а не минимизировать среднее отклонение. Кроме того, ЛБП-формулировка позволяет естественным образом включать жёсткие детерминированные ограничения, тогда как регуляризационные методы требуют для этого существенной модификации оптимизационных алгоритмов. Главным ограничением ЛБП-подхода является его более высокая вычислительная сложность по сравнению с градиентными методами. Он наиболее эффективен в задачах, где априорная информация критически важна, а размерность выборки приемлема для современных решателей смешанно-целочисленного линейного программирования. Сравнение предложенного метода с регуляризационными подходами по точности прогнозирования на тестовых выборках планируется выполнить в последующих работах.

Библиография

- [1] Draper N., Smith H. *Applied Regression Analysis*. — 3rd ed. — New York : Wiley, 1998. — 706 p.
- [2] Schmidt K. *Parameter Estimation in the Linear Regression Model with Prior Information under Inequality Constraints (In German)*. — Frankfurt am Main : Anton Hain, 1992. — 173 p.
- [3] Knopov P.S., Korkhin A.S. *Regression Analysis Under A Priori Parameter Restrictions*. — New York: Springer, 2012. — 234 p.
- [4] Slawski M., Hein M. Sparse recovery by thresholded non-negative least squares // *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*. — P. 1926–1934.
- [5] Wu L., Yang Y., Liu H. Nonnegative-lasso and application in index tracking // *Computational Statistics & Data Analysis*. — 2014. — Vol. 70. — P. 116–126.
- [6] Pya N., Wood S.N. Shape constrained additive models // *Statistics and Computing*. — 2015. — Vol. 25, No 3. — P. 543–559.
- [7] Boyd S., Vandenberghe L. *Convex Optimization*. — Cambridge : Cambridge University Press, 2004. — 716 p.
- [8] Geweke J. *Contemporary Bayesian Econometrics and Statistics*. — Hoboken, N.J. : John Wiley & Sons, 2005. — 300 p.
- [9] Носков С.И., Шахуров А.Н. Максимизация числа допустимых ошибок аппроксимации при построении линейной регрессионной модели // *Вестник Югорского государственного университета*. — 2024. — № 3. — С. 57–62.
- [10] Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimization lens // *The Annals of Statistics*. — 2016. — Vol. 44, No 2. — P. 813–852.
- [11] Носков С.И. *Технология моделирования объектов с нестабильным функционированием*. — Иркутск, 1996. — 320 с.
- [12] Носков С.И., Бычков Ю.А., Перфильева К.С. Разработка регрессионной модели пассажирооборота воздушного транспорта Российской Федерации двумя альтернативными методами // *Вестник кибернетики*. — 2023. — № 22 (1). С. 36–42.

Maximizing the number of acceptable errors in linear regression approximation with deterministic constraints

S.I. Noskov, V.M. Lacviev

Abstract — The paper considers the problem of constructing a linear regression model aimed at maximizing the number of observations for which the approximation error does not exceed a given threshold value. Unlike classical methods that minimize the sum of squares or error modules, the proposed approach allows you to directly control the proportion of acceptable deviations. The problem is formulated in the form of a linear-Boolean programming (LBP) model. In addition to the basic model, six types of deterministic constraints are proposed that reflect a priori expert knowledge about the subject area: interval constraints on parameters, constraints on the sum and maximum of absolute errors, sign compatibility of coefficients, the requirement of accurate hits (the presence of a given number of observations with zero error) and the uniformity of the contribution of features across observations. It is shown that all the introduced restrictions allow linearization, which preserves the belonging of the problem to the LBP class. The statement about the equivalence of the initial and linearized statements is formulated and proved. The proposed approach is used to model the passenger turnover of the Russian Federation's air transport based on statistical data for 2002-2019. A comparative analysis with previously published results shows that the introduction of additional restrictions leads to a tightening of the allowable set of solutions: the number of allowable errors decreases from 11 to 8, however, the sum of error modules decreases from 175.6 to 167.8, which indicates an increase in the accuracy of approximation over the entire set of observations.

Keywords — regression, linear Boolean programming, a priori information, deterministic constraints.

REFERENCES

- [1] Draper N., Smith H. Applied Regression Analysis. — 3rd ed. — New York : Wiley, 1998. — 706 p.
- [2] Schmidt K. Parameter Estimation in the Linear Regression Model with Prior Information under Inequality Constraints (In German). — Frankfurt am Main : Anton Hain, 1992. — 173 p.
- [3] Knopov P.S., Korkhin A.S. Regression Analysis Under A Priori Parameter Restrictions. — New York: Springer, 2012. — 234 p.
- [4] Slawski M., Hein M. Sparse recovery by thresholded non-negative least squares // Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. — P. 1926–1934.
- [5] Wu L., Yang Y., Liu H. Nonnegative-lasso and application in index tracking // Computational Statistics & Data Analysis. — 2014. — Vol. 70. — P. 116–126.
- [6] Pya N., Wood S.N. Shape constrained additive models // Statistics and Computing. — 2015. — Vol. 25, No 3. — P. 543–559.
- [7] Boyd S., Vandenberghe L. Convex Optimization. — Cambridge : Cambridge University Press, 2004. — 716 p.
- [8] Geweke J. Contemporary Bayesian Econometrics and Statistics. — Hoboken, N.J. : John Wiley & Sons, 2005. — 300 p.
- [9] Noskov S.I., Shakhurov A.N. Maximizing the number of acceptable approximation errors in constructing a linear regression model // Bulletin of Yugorsky State University. - 2024. — No. 3. — pp. 57-62.
- [10] Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimization lens // The Annals of Statistics. — 2016. — Vol. 44, No 2. — P. 813–852.
- [11] Noskov S.I. Technology of modeling objects with unstable functioning. Irkutsk, 1996. 320 p.
- [12] Noskov S.I., Bychkov Yu.A., Perfilieva K.S. Development of a regression model of passenger turnover of air transport in the Russian Federation using two alternative methods // Bulletin of Cybernetics. — 2023. — № 22 (1). Pp. 36-42.